European Journal of Operational Research 320 (2025) 160–174

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Decision Support

Generating sets of diverse and plausible scenarios through approximated multivariate normal distributions

Eljas Aalto [a,b,*], Tuomo Kuosa [b], Max Stucki [b]

[a] Department of Economics, Turku School of Economics, Rehtorinpellonkatu 3, Turku, 20500, Finland
[b] Futures Platform, Bulevardi 21, Helsinki, 00180, Finland

ARTICLE INFO

Keywords:
Scenarios
Strategic planning
Decomposition methods
Forecasting

ABSTRACT

This article presents a novel and broadly generalizable framework for generating diverse and plausible sets of scenarios. Potential future outcomes are decomposed using a set of uncertainties which are assumed to be multivariate normally distributed, regardless of whether the uncertainties actually present numerically quantifiable phenomena. The optimal scenarios are then chosen along the principal components of the distribution, and the results can be easily interpreted and visualized. Notably, our approach requires a relatively small number of numerical assessments, offering an efficient and practical solution for decision-makers. The framework also provides a testable setting for evaluating its performance and allows users to iteratively improve future-related assumptions and predictions. These findings are relevant for all fields that aim to understand potential future developments, such as, but not limited to, foresight, economics, business strategy and strategic intelligence analysis.

1. Introduction

Future studies and foresight have been criticized for not adhering to established scientific practices such as rigorous theory building and testing, and consequently, the attitude towards future studies is highly skeptical among other scientific fields (Fergnani & Chermack, 2021). While scenario planning stands out as a widely recognized approach stemming from the futures studies literature, Cordova-Pozo and Rouwette (2023) argue that conceptual confusion, methodological chaos and scarcity of evidence on its effectiveness limit its spread. Future uncertainty is seldom assessed based on actual probability distributions, but instead the analysis usually relies on highly subjective, qualitative judgements elicited from various domain experts (Seeve & Vilkkumaa, 2022). The emphasis on expert assessments arises both from the attempt to capture profound structural changes within long time horizons (Börjeson et al., 2006) and the reluctance of futures studies to engage in explicit forecasting (Fergnani & Chermack, 2021). However, it is well documented in the relevant literature that there exists biases in these types of expert judgements (e.g. Ahrenshop et al., 2023; Collaborative, 2023; Fildes et al., 2009; Lawrence et al., 2000).

This article tries to tackle some of these shortcomings by providing a novel, practical and transparent framework to generate exploratory scenarios. In light of the above-mentioned critique, the method has two highly valuable properties: (1) interlinkages to simple and scientifically established concepts and (2) the systematic testability of future-related assessments and assumptions that the analysis is based on. The proposed framework builds on several existing methods and techniques utilized in scenario planning and contributes to the growing literature in the domain. It shares the same goal of e.g. Seeve and Vilkkumaa (2022), Lord et al. (2016) and Tietje (2005) as it aims at creating a relatively small number of explorative and decomposed scenarios that are both maximally plausible and maximally diverse.

The explorative scenario project goal refers to the aim to explore situations or developments that are regarded as possible. They respond to the question: *what can happen?* (Börjeson et al., 2006; Kowalski et al., 2009). These types of scenarios help to explore developments that the intended target group in one way or another may have to take into consideration. In case they are used to inform strategy development of a planning entity, they for example help the decision-maker to develop robust strategies, i.e. strategies that will survive several kinds of external development (Börjeson et al., 2006). Due to this, in explorative scenario projects it is natural to aim at maximal scenario diversity and plausibility, which are commonly used criteria to assess the quality of the generated scenario set (Börjeson et al., 2006; Bunn & Salo, 1993; Seeve & Vilkkumaa, 2022). The number of scenarios should also be kept relatively small due to manageability (Heugens & van Oosterhout, 2001) while still avoiding missing any relevant possibilities (Schweizer & Kriegler, 2012).

* Corresponding author at: Department of Economics, Turku School of Economics, Rehtorinpellonkatu 3, Turku, 20500, Finland.
E-mail addresses: elkuaa@utu.fi (E. Aalto), tuomo.kuosa@futuresplatform.com (T. Kuosa), max.stucki@futuresplatform.com (M. Stucki).

https://doi.org/10.1016/j.ejor.2024.08.003
Received 1 February 2024; Accepted 1 August 2024
Available online 2 August 2024

Decomposed approaches in turn refer to analysis tools that break the target down into smaller pieces that are isolated and systematically studied further. In practical terms this means decomposing the uncertain future into a set of separate variables, called *uncertainty factors*, that are the key drivers having an effect on the topic and that have different outcome levels (Tietje, 2005; Wright et al., 1988). The scenarios are then developed based on combining these different uncertainty factor levels. By utilizing decomposition, the task can be broken down into cognitively less demanding sub-tasks (Seeve & Vilkkumaa, 2022).

There are two main classes of decomposed scenario methodologies in the domain of futures studies. *Cross-Impact Analysis* (CIA) is one of the most commonly used class of decomposition methods, and generally this umbrella term refers to the methods where probabilities for scenarios are explicitly defined. In CIA, several trends or uncertainty factors are put into a matrix to represent the dependencies and influences among the elements. An expert panel assigns probabilities or weights to indicate the likelihood and strength of the impact one variable may have on another, and these assessments are then aggregated into numerical estimates about the joint probabilities of combinations of levels (Gordon, 2009; Helmer, 1981). CIA and its derivatives have been utilized for example in Roponen and Salo (2023), Kluyver and Moskowitz (1984), Brauers and Weber (1988) and Salo et al. (2022).

Another related class of methods falls under the other umbrella term of *consistency analysis*. These methods focus on the compatibility of each factor-level pair on a more general level without explicitly relying on probabilities. Ideally, one would examine all of the level combinations in the set of scenario possibilities in order to conclude which of them are possible, viable, and logical. In doing so, one marks out a consistent subset of scenarios. Yet, due to the exponential nature of the number of possibilities in the set, manual analysis becomes impossible. Therefore, a useful step in the analysis process is to research the internal relationships with a consistency measure that assesses whether each factor-level pair can logically coexist. Consequently, a large number of mutually contradictory factor-level combinations can be discarded, and by utilizing this technique, the number of possibilities can be reduced by even 99%, depending on the problem structure (Ritchey, 2006). Examples of methods utilizing consistency analysis are (Seeve & Vilkkumaa, 2022), Carlsen et al. (2016), Schweizer and Kriegler (2012) and Tietje (2005).

However, there are several challenges related to both CIA and consistency analysis. Firstly, the discretization of factor levels might be relatively arbitrary. As a result, scenario probabilities are dependent on the number and granularity of uncertainty factor and outcome levels, and a single scenario probability does not communicate any useful information in itself. The more granular the scenarios are, the lower the probabilities.

Secondly, both methods often still require a very large number of numerical assessments as the number of separate factor-level pairs grows rapidly by increasing the number of uncertainty factors or levels. This might lead to impracticality and a situation where important factors are too easily omitted (Brauers & Weber, 1988; Lord et al., 2016).

Thirdly, although the set of consistent or sufficiently probable scenarios can often be narrowed down, there exists much ambiguity related to the selection of the final set of scenarios. Different methods for final scenario selection have been proposed, such as clustering tools (Brauers & Weber, 1988), integer linear programming (Jenkins, 1997; Seeve & Vilkkumaa, 2022) and iterative selection methods (Tietje, 2005), but no widely established criteria exist.

Fourthly, the optimization algorithms that aim at finding the optimally diverse and plausible set of scenarios might be computationally intensive (Roponen & Salo, 2023). This restricts the use of highly complex methodologies in practical settings.

Lastly, the analysis often relies solely on subjective judgement, the ability of which is questionable and which is seldom explicitly evaluated *ex posteriori* in the context of scenario analysis. For example, Ahrenshop et al. (2023) find that caution should be exercised especially when undertaking expert forecasting, since experts may have unrealistic expectations and may be inflexible in altering these even when provided new information. These types of biases may lead to suboptimal results, and the added value of highly complex methodologies becomes questionable. Also common Delphi-like iterative approaches or metrics such as intercoder reliability might be prone to collective biases.

The novel method presented in this article closely resembles CIA and consistency analysis. However, the novelty of the method follows from the feature that uncertainties are modeled and interpreted as continuous and standard Gaussian variables, leading to an infinite set of different level combinations between the variables. The normality assumption has a long history of successful application in various fields, from finance and engineering to the natural and social sciences. Many natural processes involve the combined effects of numerous independent factors. When these factors contribute to a phenomenon, the resulting distribution may approximate a normal distribution (e.g. Lyon, 2014). This is observed in areas such as measurement errors, biological variability and financial markets. Thus, normality can also be a useful approximation for the behavior of uncertainty factors.

In our novel method, the relationships and co-dependencies between uncertainties are assessed *a priori* by using linear correlation scores that can either be based on actual data or expert judgement. Assuming the dependency between variables is linear is the most common way in social sciences to represent relationships between different phenomena, and it is also the core assumption behind linear regression analysis. Its simplicity, interpretability, and wide applicability make it a go-to method for initial analyses in various scientific fields. This also simplifies the task of internal scenario consistency assessment by significantly narrowing down the number of numerical co-dependency assessments relative to methods where the uncertainty factor levels are discrete. This also adds a level of generality, as quantitative variables and empirical estimates can easily be combined with more qualitative assessments. Thus, the analysis does not necessarily rely solely on expert judgement.

In other words, we approximate an actual multivariate normal probability distribution for the future state, and the optimal scenario set is picked from the approximate distribution by using some formal optimization criteria. In this paper we define that optimality is based on the diversity and plausibility of the selected scenarios, and by utilizing this criterion, we show that the optimal set of scenarios is obtained by picking the scenarios along the principal components of the approximated distribution. The same idea has been utilized also to generate numerical risk scenarios on the financial markets (e.g. Loretan, 1997; Novosyolov & Satchkov, 2008), and also Seeve and Vilkkumaa (2022) use an analogue for principal component analysis that is suitable for categorical data – multiple correspondence analysis – in their framework. The scenario set optimization using principal components does not require computationally intensive optimization algorithms as the results can be obtained by simply using the theoretical distribution. The central focus in the analysis should lie on the assessment of co-dependencies and the precise interpretation of the continuous numerical scale (see Section 2).

Another novelty is the possibility of *ex posteriori* evaluation. This evaluation does not focus on the effects of the practical implementation of the scenarios, but instead on the future-related *assumptions* that the choices in the analysis process are based on. Because scenario planning methods usually rely on subjective judgements, it is crucial to be able to systematically test the assumptions and perceptions that the judgements are based on, especially as the biases in expert forecasts are well-known. In addition, the assumptions themselves can be elusive and hard to detect, even to their holders. As Coates (1999) points out, getting to assumptions that underpin future-oriented analysis can be very troublesome — and certainly will not become easier after the

scenarios have been made and are subject to a post mortem, which is a significant problem for the aim of increasing the plausibility and forecasting accuracy of scenario analysis. The proposed method not only makes it possible to evaluate and correct the assumptions at play, as mentioned above, but it also makes them visible from the very start of the analysis and documents them for further scrutiny.

The testing requires gathering a set of relatively independent forecasts and observations regarding different uncertainty factors so that the distributional assumptions can be tested after a desired time period (see Section 4). Consequently, systematic biases in future assumptions can be eliminated and the method and its applicability can be improved iteratively. This can, however, take several years, as scenarios often involve time periods many years in the future.

The remainder of the article is structured as follows. After the introduction, we present how the distributional approximations are formed, and in the third section we show how scenarios are optimally selected based on these approximations to maximize the scenarios' diversity and plausibility. In the fourth section we present how the performance of the framework and its users' assumptions can be evaluated and iterated. The fifth section provides a case example regarding the electric car market in 2030, and section six concludes.

2. The uncertainty distribution

2.1. Setup

Let us assume we are building scenarios describing the state of the future of some topic at time T.[1] As with other decomposition methods, the future state is determined by the outcome of a set of N different uncertainty factors related to and relevant for the topic. In this framework we assume these factors are *external*, i.e. not controllable by the actor. The most suitable N depends on the context, but it has to cover all significant trends, change drivers, change blockers, events and other external variables which can have a significant effect on the topic under analysis. For the sake of concreteness, a suitable number could vary from 5 to 10. Hereinafter, these factors will simply be called *uncertainties*.[2]

In futures studies and forecasting, there are several established heuristics for identifying the most critical uncertainties. These include e.g. *Force Field Analysis* (Lewin, 1951), which divides the situation into sets of driving forces and restraining forces, *Futures Triangle* (Inayatullah, 2008) that considers the push of the present, the weight of the past, and the pull of the future, and *PESTLE* (Aguilar, 1967), which helps to identify driving forces related to political, economic, social, technological, legal, and environmental phenomena. There are also many other types of participatory polls and qualitative frameworks. A common denominator for all methods is some type of expert knowledge. The final uncertainty selections must naturally be tailored based on the specific needs, goals and context.

In our setup, the uncertainties must have a representation on a continuous *cardinal* scale from minimum to maximum. This does not imply that all nominal uncertainties must be excluded, because most nominal uncertainties can also be represented on a continuous and cardinal scale. As an example, the nominal uncertainty *the winner of the parliamentary election* could often also be represented by the variable *the support of the left/right/party x*. Contrary to traditional decomposition methods that use a set of uncertainties, we assume that the uncertainties can take any value between the two extremes, and the interval is therefore not divided into a discrete set of separate levels

The uncertainties themselves do not need to be numerical: qualitatively expressed uncertainties are equally useful, as long as they have some type of direction guaranteeing that their outcome can be interpreted on a scale from minimum to maximum. As an intuitive example, the tightness of regulation might not be easily measured numerically, but the interpretation on the scale from minimum to maximum is still relatively easy to understand. In fact, we assess that in the majority of practical applications, uncertainties with directly usable numeric indicators are a minority. Binary variables, i.e. events that happen or do not happen, can also be used. In this binary case, the continuous scale can be interpreted to represent the probability for the event, and when the probability surpasses 50 percent, we can assume that the event happens.

The development level of each uncertainty at time T is represented by a random and continuous Gaussian[3] variable $U_i \in \mathbb{R}$, where $i \in \{1, \dots, N\}$. As was explained earlier, the normality assumption has a long history of successful application in various fields, and thus it is a useful approximation for the behavior of uncertainties. We assume that $\mathbb{E}(U_i) = \mu_i = 0$ for all $i \in \{1, \dots, N\}$ which implies that the level $U_i = 0$ represents the *expected* outcome for each uncertainty. In other words, zero represents the center of the potential outcome distribution which should not be confused with the center of the overall *range* for possible outcomes. The value zero could often be understood as a business-as-usual trend level, but more generally its interpretation is the answer to the question: what is the most probable outcome for the specific uncertainty in the selected time frame?

As an intuitive example, fertility (measured by births per woman) is a critical uncertainty related to many different topics. If the developed world was the context of the scenarios, the zero mean could be set to represent something around 1.5 births per woman, depending on the specific country and time frame. It is also important to note that in this case the value $U_i = 0$ would represent significantly lower fertility rates compared to the case of the developing world.

We scale and interpret the variables so that for each U_i, the value $U_i = -2$ represents the *minimal* development level, whereas a value of $U_i = 2$ represents the *maximal* development level. This gives us fixed reference points for the interpretation of the numerical scale. Although the scale from minimum to maximum is conceptually easy to understand, it is crucial to clearly define its practical interpretation. Because the random variable itself is not truncated and it can also take values outside the interval $[-2, 2]$, the two opposite thresholds are interpreted to represent two opposite, equally (un)likely and extreme development levels that are equally distant from the expected value. This interpretation of the scale ensures that the variance of each variable U_i can be set equal, and the probability of each variable surpassing the threshold levels -2 and 2 is similar.

We denote $\mathrm{Var}(U_i) = \sigma_i^2 = \sigma^2$ for all $i \in \{1, \dots, N\}$. A natural benchmark level for the variance is $\sigma^2 = 1$ so that each variable follows a standard normal distribution. When $\mu_i = 0$, this implies that the thresholds -2 and 2 are two standard deviations away from the mean, and approximately 5 percent of random outcomes would fall outside the interval. The levels -2 and 2 should therefore be interpreted as extreme development levels that in total have a 5 percent probability of occurrence. The actual meaning and interpretation of the thresholds should be explicitly described based on this intuition.

[1] Because the time frame for the scenarios is fixed, we do not use time indices in the notation.

[2] This is due to the fact that factor variables often refer to categorical variables, and in our analysis the variables are assumed to be continuous.

[3] More generally, the uncertainties can also be represented by any other random distribution. In this paper, we assume the uncertainty distribution is always normal due to its wide applicability and useful properties. To be more exact, the probability for binary events should be derived from the cumulative distribution function such as in the case of probit models. Moreover, for time series variables the possible autoregression and the order of integration should also be taken into consideration, However, as the framework is only a rough and qualitative approximation of the behavior of uncertainties which should be kept efficient and practical, this might not be necessary.

The assumption $\sigma_i^2 = 1$ is especially suitable, because data points deviating more than two standard deviations from the mean are typically considered outliers in a normally distributed data. In futures studies a 5 percent probability could be intuitively used as an approximate level for *wild card* (Petersen & Steinmueller, 2009) or *black swan* (Taleb, 2007) events that are established terms to describe low-probability, high-impact outlier outcomes. Thus, the thresholds have an interpretation that is closely linked to established concepts both in futures studies and the natural sciences. However, although $\sigma_i^2 = 1$ is a natural choice, it is ultimately up to the analyst and the specific context how the standard deviation is chosen.

It is important to note that the sign of the variable U_i does not indicate the sign of the uncertainty itself: the scale from -2 to 2 is an arbitrary cardinal representation for the interval from minimum to maximum, and it should not be confused with the actual level of uncertainty outcomes. It does not indicate e.g. whether the growth rate of some specific phenomenon used as an uncertainty is positive or negative. Moreover, the scale is not a ratio scale: as an example, the value 2 for economic growth does not imply double growth rate compared to 1.

Based on the definition of the variables and their distribution, the future state can be represented by a random N-dimensional multinormal variable $\mathbf{U} = (U_1, \dots, U_N)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean vector $\boldsymbol{\mu} = \mathbf{0}$ is an N-dimensional vector of zeros and $\boldsymbol{\Sigma}$ is the covariance matrix. A single data point $\mathbf{u} = (u_1, \dots, u_N)^T$ characterizes one future outcome - a *scenario* - and it can be qualitatively interpreted and described using a narrative. We formalize this definition as follows:

Definition 1. A scenario is an N-dimensional vector $\mathbf{u} = (u_1, \dots, u_N)^T \in \mathbb{R}^N$.

The mean vector $\boldsymbol{\mu}$ describes the expected state of the future, or the *expected scenario*. The qualitative interpretation of each scenario data point naturally depends on the precise definition of the interval $[-2, 2]$ and our assessments about the expected outcomes.

After choosing the suitable set of uncertainties and assessing their scale, it is necessary to qualitatively assess or numerically estimate the covariance matrix $\boldsymbol{\Sigma}$. Because all individual variables have unit variance, the covariance matrix equals the correlation matrix. It is defined as the following $N \times N$ matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{N1} & \rho_{N2} & \cdots & \rho_{NN} \end{pmatrix}, \tag{1}$$

where ρ_{ij} denotes the correlation coefficient between uncertainties i and j. Naturally, $\rho_{ij} = 1$ when $i = j$, and $\rho_{ij} = \rho_{ji}$.

Hence to be able to form the distribution, we still need to form the correlation matrix. In the next section, we present how the co-dependencies between uncertainties are assessed using correlations.

2.2. Co-dependency assessment by correlations

Generating a useful and plausible set of scenarios requires assessing the co-dependencies between different uncertainties. Consequently, we do not assume the uncertainties are independently distributed, i.e. we allow the outcomes of uncertainties to be dependent on or influenced by the outcomes of the other uncertainties.

With categorical uncertainty levels, pairwise consistency assessments can be used to find a useful subset of possible scenarios. These pairwise consistency scores indicate whether each uncertainty level pair is logical and plausible, and the final scenarios are usually formed by choosing the most diverse and consistent combinations. In our case the continuity of the random variables makes consistency analysis impossible as there is a continuum of different combinations. However, the random variables can be correlated.

We simplify the dependency between each variable pair by assuming a linear relationship between them. Thus, Pearson's correlation coefficient varying between -1 and 1 can be used as a natural analogy to consistency scores. The stronger the correlation, the more likely it is that the direction of development of two uncertainties is identical or reverse, depending on the sign. As an example, economic growth rates and unemployment have historically had a strong negative correlation, and this relationship is known as Okun's law (Prachowny, 1993). Correlations can be used regardless of whether some uncertainties represent level variables, some growth rates, and some binary variables. The correlation can also be assessed regardless of whether the uncertainties represent concepts that can easily be quantified.

The use of correlation coefficients as co-dependency measures can be motivated for two main reasons. First, it requires much fewer co-dependency assessments. The number of correlations to be assessed is equal to $(N^2 - N)/2$, whereas with categorical uncertainties and L different levels, the number of consistency estimates is L^2 times larger. In categorical consistency analysis the number of different level combinations also grows rapidly by increasing the number of uncertainties and their different levels. For instance, with 10 uncertainties and 4 levels, the required number of categorical assessments is 720, whereas by using continuous variables the number of correlations is only 45. This means that important uncertainties do not need to be discarded in the analysis due to time or other resource constraints.

Secondly, assuming the dependency between variables is linear is the most common way in social sciences to represent relationships between different phenomena. Linear relationships are well-studied and understood in statistics and mathematics, allowing for the application of a wide range of established analytical techniques and tools such as linear regression. While linear relationships may not capture the full complexity of real-world phenomena, they serve as a valuable starting point for analysis and hypothesis testing.

Assessing correlations does not require identifying causal relationships: a strong correlation does not tell whether uncertainty A causes uncertainty B, whether uncertainty B causes uncertainty A, or whether they have any causal relationship at all. It only represents a statistical dependency. The correlations can be assessed with an accuracy of one decimal and filled in a matrix as shown by Eq. (1). As with the initial identification of uncertainties, this can be conducted easily in a workshop setting. Also Delphi-like processes can be used to estimate the correlations so that numerical responses are aggregated from a larger group of experts.[4] Optimally, there are data and estimates directly available, but we assess that such cases are a minority. However, this is largely dependent on the context, as with e.g. economic and financial variables these cases might even be a majority. Most importantly, it is the intuition about the *sign* and the approximate *strength* of the dependency that have to be assessed.

There are some caveats related to the use of correlations. First, there might be confusion over the concept of correlation and its definition if the practitioners are not familiar with statistical analysis. Causal analysis might often be performed "one-way", i.e. only the effect of A on B is assessed, and the effect of B on A is forgotten. Moreover, third factors that might have an effect on two seemingly unrelated uncertainties can be forgotten, and these correlations are consequently underestimated. Thus, the process clearly requires a very well-explained definition for the statistical concept.

Secondly, assessing correlations between variables and phenomena that have clear nonlinear dependencies will cause bias. A good example is the inverted-U relationship between competition and innovation, observed in economic research (Aghion et al., 2005). In this case, the

[4] For a classical paper on Delphi-based approaches, see Rowe and Wright (1999). Barrios et al. (2021) and O'Hagan (2019) provide more recent analyses.

uncertainties might have a near-zero correlation, which totally hides their strong co-dependency.

Thirdly, we believe that without data, correlations are often overestimated due to overconfidence and they become unrealistically strong. Overestimation also often leads to inconsistencies in the correlation table and consequently some of its eigenvalues might become negative. Thus it should be made clear that strong correlations in social sciences are a rarity. One way to overcome this problem is to give minimum and maximum correlation thresholds of e.g. -0.5 and 0.5 which cannot be exceeded. Qualitative assessments could also be used, and these qualitative levels should have a realistic and consistent numerical coding (e.g. an "extremely high positive correlation" could be coded as 0.5).

Fourthly, correlations should represent the uncertainties' relationship in the future. Hence, relying only on information about the historical dependency can cause bias if the factors effecting the variables' relationship can be assumed to differ from the past. A classic example of this is the pre-1970 Keynesian assumption about the negative relationship between inflation and unemployment, represented by the Phillip's curve (Phillips, 1958). The first oil shock and the following stagflation caused the global economy to suffer from both high inflation rates and high unemployment, leading to the old theory becoming obsolete. Another time-related problem might be the difference between long-run and short-run correlations. As an example, money supply and inflation have a near-zero short-run correlation, but they might have a strong positive correlation in the long run (Gertler & Hofmann, 2018). Thus, the time frame must be taken into account very carefully.

After the correlation table has been formed, we have defined all the parameters characterizing the distribution $\mathbf{U} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We call this distribution the *approximated distribution*. The next step is to generate an optimal set of scenarios based on this distribution.

3. Forming the optimal set of scenarios

After forming the approximate distribution for the relevant uncertainties, we need to select the optimal set of scenarios. As was presented, a *scenario* is defined as a vector of uncertainty outcomes, and scenario j is denoted with $\mathbf{u}_j = (u_{j1}, \dots, u_{jN})^T$ for $j \in \{1, \dots, M\}$ where M is the total number of scenarios. We denote the set of all scenarios by $S = \{\mathbf{u}_1, \dots, \mathbf{u}_M\}$.

Exploratory scenario analysis often means optimization between three central features: on the one hand the *number* of scenarios must be minimized, but on the other hand the scenarios must be sufficiently *diverse* to cover a maximal range of future possibilities. Additionally, the scenarios must be *plausible* which in our case refers to their probability and internal consistency. In conclusion, exploratory scenario analysis has to cover the future possibilities in a small number of maximally diverse and plausible scenarios (Börjeson et al., 2006). In this section, we show that with a specific intuitive definition for the diversity, plausibility and optimality of a scenario set, the scenarios are selected along the principal components of the approximated distribution.

3.1. Measuring scenario diversity and plausibility

Scenario differences are intuitively measured using the *distance* between the chosen data points (Lord et al., 2016; Tietje, 2005). The larger the distances between the scenarios are, the larger their differences. The most natural option to measure distance is by using Euclidean distances, but more generally the distance between any data point $\mathbf{u}_j$ and $\mathbf{u}_k$ is given by function $D(\mathbf{u}_j, \mathbf{u}_k)$. Hereinafter distance refers to the Euclidean distance, and thus we define that

$$D_E(\mathbf{u}_j, \mathbf{u}_k) = \sqrt{\sum_{i=1}^{N} (u_{ji} - u_{ki})^2}\,, \tag{2}$$

where u_{ji} denotes the i:th element in the j:th scenario. The lower index E refers to Euclidean.

The distance function D_E only measures diversity between two scenarios. How then should the diversity of the whole set S be measured? We propose that the diversity of S is denoted by function $G_d(S)$ and it is given by the minimum pairwise distance in the set, i.e.

$$G_d(S) = \min_{\substack{\mathbf{u}_j, \mathbf{u}_k \in S \\ j \neq k}} D_E(\mathbf{u}_j, \mathbf{u}_k)\,. \tag{3}$$

Other options for function $G_d(S)$ include e.g. total variance and average pairwise distance. However, contrary to several other options the minimum pairwise distance ensures that maximizing diversity does not lead to any two scenarios becoming too similar. Another diversity measure that is relatively sensitive to single short distances in the set is the harmonic mean that is used by Tietje (2005) in the so-called Distance-To-Selected (DTS) scenario selection procedure. However, the application of this measure in the continuous distribution presented in this paper is left for future research.

When it comes to scenario plausibility, in a continuous uncertainty space the number of possible future outcomes is infinite and respectively the probability of each separate state is zero. Additionally, we do not have any consistency measures for the different outcomes. Because we are interested in assessing scenario plausibility, we clearly need some other measure than explicit probabilities or consistency scores. The probability density is the most natural choice, but we propose another measure due to its useful properties and simplicity. This measure is the *Mahalanobis distance* which measures the distance between a point and a distribution. It takes into account the correlations between different variables, and it is a way of measuring how many standard deviations away an observation is from the mean, considering the shape of the distribution. Its value is closely related to the probability density.

Formally, the Mahalanobis distance is given by:

$$D_M(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{(\mathbf{u} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{u} - \boldsymbol{\mu})}\,, \tag{4}$$

where $\mathbf{u}$ is the vector representing the scenario, $\boldsymbol{\mu}$ is the mean vector of the distribution and $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix. In our case the mean vector is just a vector of zeros, and thus we drop $\boldsymbol{\mu}$ from the equation and it simplifies to

$$D_M(\mathbf{u}; \boldsymbol{\Sigma}) = \sqrt{\mathbf{u}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}}\,. \tag{5}$$

We define that the plausibility of the whole set S is measured by function $G_p(S; \boldsymbol{\Sigma})$. Contrary to function G_d, the measure of plausibility naturally also depends on the correlation matrix $\boldsymbol{\Sigma}$. We propose that the plausibility of the set is given by the negative of the maximum squared Mahalanobis distance in the set S, i.e.

$$G_p(S; \boldsymbol{\Sigma}) = -\max_{\mathbf{u}_j \in S} (D_M^2(\mathbf{u}_j; \boldsymbol{\Sigma}))\,. \tag{6}$$

In this way, the plausibility of the set decreases the more the strongest outlier deviates from the mean. The maximum function will effectively prevent any scenarios from becoming too extreme, and the squaring will more strongly penalize large Mahalanobis distances. In addition, the squaring will later ensure that the scenario optimization problem has an interior solution. The negative of the average Mahalanobis distance or the harmonic mean could be another good (and even more intuitive) option for the function $G_p(S)$. However, in the following we will use the squared maximum for analytical simplicity, and the other options are again left for future research.

Following the idea of the DTS procedure by Tietje (2005), the function that measures both desired properties in the set S can be formally defined as an objective function $P(S; \boldsymbol{\Sigma})$ that is given in a standardized form as follows:

$$P(S; \boldsymbol{\Sigma}) = \alpha_d G_d(S) + \alpha_p G_p(S; \boldsymbol{\Sigma})\,. \tag{7}$$

The parameters $\alpha_d, \alpha_p \geq 0$ are weighing parameters that determine the weight of each separate feature in the function. If $\alpha_d = 0$, we are only interested in maximal scenario plausibility. Similarly when $\alpha_p = 0$, we are only interested in maximal scenario diversity.

When we are interested in both properties, only the ratio α_d/α_p has an effect on the optimum, not the parameters' absolute value. Thus, we can normalize $\alpha_p = 1$, and consequently only the weight of diversity has to be adjusted.

3.2. Solving the optimum

Now when we have defined how diversity and plausibility in the set S are measured, we can proceed to the actual optimization. First, it is natural to restrict the possible scenario data points by setting the thresholds -2 and 2 as the lowest and highest possible values for each individual element in each scenario vector $\mathbf{u}_j$. Thus the scenarios are constrained by

$$-2 \le u_{ji} \le 2 \text{ for all } j \in \{1, \dots, M\} \text{ and } i \in \{1, \dots, N\}, \tag{8}$$

which implies an N-dimensional hypercube for the possible scenarios.

We propose that the number of scenarios M is either 2, 4 or 6 which is approximately in line with the suggestion by Lord et al. (2016). Moreover, according to Miller (1956), the number 6 is an upper limit for manageability. The odd numbers are excluded because later it turns out they do not produce unique optima. Moreover, with an odd number the scenario set does not become balanced, i.e. its centroid does not lie in the origin.

We propose that the final number (2, 4 or 6) is selected based on the eigenvalues of the covariance matrix and their ratios. Eigenvalues describe variances along the principal components which are the directions along which the distribution has the largest variance. When there is a large number of directions with high variance, it is natural that a larger number of scenarios is required. Conversely, when there are only a few directions along which the distribution has a large variance, the number of scenarios is smaller. When the total variance in the uncertainty space is sufficiently explained only by one principal component, the number of scenarios is 2. Similarly, when two or three principal component explain a sufficient share of variance, the number of scenarios is 4 or 6 respectively. However, we do not want that the number exceeds 6 and thus it is set as maximum for M, although there would be additional principal components with relatively high variance.

Let $1/3 \le c < 1$ be some threshold value. We propose that if the eigenvalues $\lambda_1, \dots, \lambda_N$ are given in a descending order so that $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N$, the number of scenarios is selected as follows:

If $N = 2$:

$$M = \begin{cases} 2, & \text{if } \lambda_2/\lambda_1 < c \\ 4, & \text{otherwise} \end{cases} \tag{9}$$

If $N > 2$:

$$M = \begin{cases} 2, & \text{if } \lambda_2/\lambda_1 < c \\ 4, & \text{if } \lambda_2/\lambda_1 \ge c \text{ and } \lambda_3/\lambda_1 < c \\ 6, & \text{otherwise} \end{cases} \tag{10}$$

The value of c controls how high the variance of additional principal components must be relative to the first principal component for them to be used. This definition ensures that the total number of scenarios is larger when the number of directions with higher variance is larger. The definition also ensures that the total number of scenarios does not exceed 6. A good choice for c could be e.g. 1/2 or 2/3.

There are also several other heuristics that could be used in choosing the number of principal components. These include e.g. (1) ignoring principal components at the point at which the next one offers little increase in the total explained variation, (2) including all principal components up to a predetermined threshold of explained variation, (3) ignoring components whose explained variation is less than 1, which means these principal components offer less than one variable's worth of information (Kaiser criterion), and (4) ignoring the last principal components which explain a roughly equal share of variation. However, the best choice will ultimately depend on the context.

Now the optimization can be formally defined as maximization of the objective function (7) with respect to the vectors $\mathbf{u}_1, \dots, \mathbf{u}_M$ subject to constraint (8). This leads to the following definition for the optimal scenario set, denoted by S^*:

Definition 2. Given the covariance matrix $\mathbf{\Sigma}$, the optimal scenario set S^* is the set that maximizes function (7) subject to (8) and where M is given by Eqs. (9) and (10).

How should the M scenarios lie in the variable space, given the objective function? Based on the definition for the optimum, we obtain one of our main results: the optimal scenarios are generated by selecting two scenario data points from the opposite ends of each one of the selected principal components. The norms of the scenarios are proportional to the square roots of the eigenvalue of their corresponding principal component. The more diversity is emphasized, the farther away from the origin the scenarios should lie. The whole process is computationally light, as only the calculation of eigenvalues and their corresponding eigenvectors is required.

To present the formal result, we will first transform the scenarios into the principal component space. Let W be the matrix whose columns are the eigenvectors of the correlation matrix $\mathbf{\Sigma}$. Then the transformed scenario vector $\mathbf{z}_j^T = (z_{j1}, \dots, z_{jN})$ in the principal component space, where the individual elements are the values of the principal components, is given by $\mathbf{z}_j^T = W^T \mathbf{u}_j$. Because the optimal scenarios lie on the principal components, the transformed vector will have a single non-zero component, which corresponds to the principal component on which the scenario lies.

In the following we normalize $\alpha_p = 1$. In the optimum we have that the individual elements of the transformed vectors $\mathbf{z}_j^T = (z_{j1}, \dots, z_{jN})$, where $j \in \{1, \dots, M\}$, are given as follows:

If $M = 2$:

$$z_{ji} = \begin{cases} \alpha_d \lambda_1, & \text{if } j = 1 \text{ and } i = 1 \\ -\alpha_d \lambda_1, & \text{if } j = 2 \text{ and } i = 1 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

If $M = 4$:

$$z_{ji} = \begin{cases} \frac{\alpha_d}{2}\sqrt{\lambda_1(\lambda_1+\lambda_2)}, & \text{if } j = 1 \text{ and } i = 1 \\ -\frac{\alpha_d}{2}\sqrt{\lambda_1(\lambda_1+\lambda_2)}, & \text{if } j = 2 \text{ and } i = 1 \\ \frac{\alpha_d}{2}\sqrt{\lambda_2(\lambda_1+\lambda_2)}, & \text{if } j = 3 \text{ and } i = 2 \\ -\frac{\alpha_d}{2}\sqrt{\lambda_2(\lambda_1+\lambda_2)}, & \text{if } j = 4 \text{ and } i = 2 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

If $M = 6$:

$$z_{ji} = \begin{cases} \frac{\alpha_d}{2}\sqrt{\lambda_1(\lambda_2+\lambda_3)}, & \text{if } j = 1 \text{ and } i = 1 \\ -\frac{\alpha_d}{2}\sqrt{\lambda_1(\lambda_2+\lambda_3)}, & \text{if } j = 2 \text{ and } i = 1 \\ \frac{\alpha_d}{2}\sqrt{\lambda_2(\lambda_2+\lambda_3)}, & \text{if } j = 3 \text{ and } i = 2 \\ -\frac{\alpha_d}{2}\sqrt{\lambda_2(\lambda_2+\lambda_3)}, & \text{if } j = 4 \text{ and } i = 2 \\ \frac{\alpha_d}{2}\sqrt{\lambda_3(\lambda_2+\lambda_3)}, & \text{if } j = 5 \text{ and } i = 3 \\ -\frac{\alpha_d}{2}\sqrt{\lambda_3(\lambda_2+\lambda_3)}, & \text{if } j = 6 \text{ and } i = 3 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

When the values of principal components are solved, the scenarios are obtained in the original space by

$$\mathbf{u}_j^T = \mathbf{z}_j W^T, \tag{14}$$

where $j \in \{1, \dots, M\}$. Now we have the following formal result regarding the optimal scenario set S^*:

Result 1. The scenarios in the optimal scenario set $S^* = \{\mathbf{u}_1, \dots, \mathbf{u}_M\}$ are given by Eq. (14), where the individual elements in vectors $\mathbf{z}_j = (z_{j1}, \dots, z_{jN})$ are given by Eqs. (11)–(13).

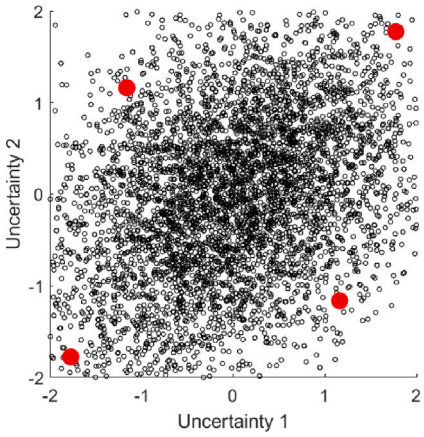

Fig. 1. $\alpha_d = 3$.

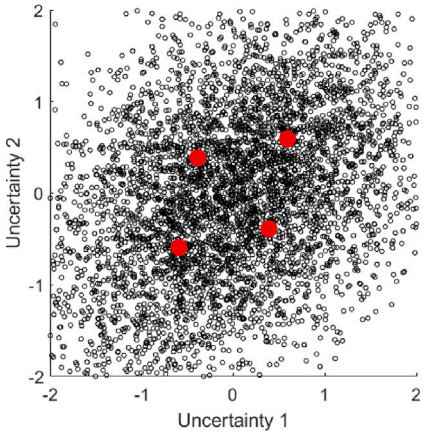

Fig. 2. $\alpha_d = 1$.

The proof of this result is provided in the appendix. It is easy to notice that increasing the weight of scenario diversity implies making the scenarios more extreme. Conversely, increasing the weight of plausibility implies moving the optimal scenarios towards the mean. In all cases $M = 2$, $M = 4$ and $M = 6$ the centroid of the scenario set is the mean of the uncertainty distribution. In other words, the centroid matches the outcome which we expect to be the most probable, which is an additional useful property. This means that the resulting scenarios are not biased towards any specific direction in the variable space.

The effect of the weights in the objective function are visualized in Figs. 1 and 2. We assume a 2-dimensional case where the correlation between the two uncertainties is 0.4. In Fig. 1 the weight on scenario diversity is larger than the weight on plausibility: consequently the scenarios, marked by red circles, lie farther away from the center. In Fig. 2 the weights of both diversity and plausibility are equal, and thus the scenarios lie closer to the origin. We have also simulated 5000 random observations from the approximate distribution for visualization purposes.

4. Evaluating assumptions and performance

Although the optimal scenario set is a result of rigorous optimization, the novel properties in the process clearly offer no added value at all if our assumptions regarding the uncertainty distribution are wrong. Because an approximated distribution is the center of our analysis, it is crucial to be able to test our assumptions regarding its structure. We naturally only observe one outcome for each approximated multivariate distribution on a single time period, and this makes statistical testing very challenging. Thus we have to use alternative methods. We openly admit that these alternative methods are not as reliable as would be desirable, but they are a step in the right direction.

We can also test our overall ability to forecast outcomes compared to random guessing. It is important to note that the tests do not only provide information about the forecasting skill of individuals and organizations: the results might also indicate the performance of the method itself and whether it is suitable for decomposing future outcomes. If the performance of the multinormal approximation is consistently poor compared to pure randomization, the utility of the framework is questionable. However, first we must define how the data used for testing is gathered.

4.1. Gathering test data

We call a set of approximated distributions and their actual outcomes as the *test data*. This test data consists of Q different topics for which the uncertainty distribution is approximated, and similarly Q vectors describing the actual outcome. This data can be formed regardless of whether the actual scenario analysis and narrative generation is performed.

In other words, we first approximate a multinormal distribution for each multidimensional outcome $\mathbf{U}_k$, where $k \in \{1, \dots, Q\}$ is the index for the topic. The time frame for all distributions should be constant, and it could be something e.g. from 2 to 5 years. The dimensionality, i.e. the number of uncertainties, of distribution k is denoted by N_k. However, the individual uncertainties in the topics may overlap. The process includes assessing the Q correlation matrices, and defining a sufficiently precise qualitative interpretation for three separate values for each individual uncertainty: -2, 0 and 2, which correspond to the minimal wild card level, the expected level and the maximal wild card level.

The number of distributions should be large enough to gain statistical power in the later steps. In practice this number could be something around 30. With this total number of topics and with e.g. 8 uncertainties for each topic the total number of uncertainties can be up to 240, and the number of individual correlations can be up to 840. The process could be constructed for different teams or organizations with varying time frames separately to compare their relative performance and control for background factors along the lines of e.g. Tetlock (2005), Tetlock and Gardner (2015), Tetlock et al. (2014).

At the chosen time period the actual outcomes are documented. This implies forming a vector $\mathbf{u}_k^A$, where upper index A refers to *actual*, for each topic $k \in \{1, \dots, Q\}$ separately by interpreting the outcome of each individual uncertainty on a scale from -2 to 2: was the development stronger or weaker than the expectation? How close was the outcome to the minimum and maximum wild card outcomes? Especially due to this step it is necessary to define the points -2, 0 and 2 very carefully and in sufficient detail so that the precise interpretation and placement on the numerical scale is possible. Finally we have a set of Q vectors $\mathbf{u}_k^A$ and Q correlation matrices which will be used for different tests and evaluations.

However, a problem is that the test data should consist of independent observations so that the later statistical tests are unbiased. By definition, the outcomes of different uncertainties are allowed to depend on each other, which implies that the independence assumption of several statistical tests is violated. Thus, the test sample should only include uncertainties for which near-zero correlation can be reasonably assumed. This requires a separate identification process to form a sample of uncertainty outcomes that can be considered independent. If the correlation between two uncertainties is high (e.g. outside the interval $[-0.3, 0.3]$, which is a commonly used cut-off point in factor analysis), one of them should be excluded from the subset of testable outcomes. On the other hand, the actual outcomes might reveal co-dependencies that were not taken into account in the original correlation assessments.

4.2. Testing the parameter assumptions

We propose three different tests for the parameter assumptions: (1) the observed mean of uncertainties, (2) the observed standard deviation of uncertainties and (3) the number of sign reversals. These measures have different purposes: the first option measures whether our expectations are unbiased, the second measures how well we assess the probability for extreme events and the third measures our ability to assess co-dependencies. It should be noted that these tests allow adjustment of future assumptions only on a general level, and the systematic biases cannot be traced back to individual uncertainties. This is due to the fact that a single observed difference between an actual outcome and the expected (assumed) outcome does not directly indicate an error in the distributional assumption as randomness belongs to the process. Only when the errors are systematic and consistent over several uncertainties, iteration of assumptions is possible.

(1) *Expected outcomes*. If our assumptions about normality and expected outcomes hold, the observed outcomes should be normally distributed and the mean of uncertainty outcomes should approach zero. This can be tested by using the uncertainty data and conventional statistical tests such as the t-test. The result might also indicate whether our assumption about the expected outcome systemically over- or underestimates the real uncertainty outcomes. When the observed mean significantly differs from zero, information about the direction and magnitude of the systematic error is crucial for the ability to calibrate future expectations.

However, a large problem is that interpreting the nature of the systematic error might be problematic due to the diverse nature of uncertainties and the consequent problems regarding the meaning of the words over- and underestimation. As an example, if the future *support for the political left* is overestimated, the future *support for the political right* is underestimated. With binary (event) variables the problem becomes even more apparent. Also the selected set of uncertainties in the case example (Section 5) demonstrates that the term "overestimation" has no clear and interpretable general meaning regarding future assumptions.

Thus, the selection of observations in the test data should not only focus on the independence of observations, but also on the selection of uncertainties that provide any reasonable information about systematic biases. This might require dividing the uncertainties into distinct and separately testable sets that provide reasonable information about biases. In other words, the aim is to form subsets of observations that have a relatively similar interpretation for the possible bias. As an example, a separate set could be formed for the development of different future technologies so that the results indicate whether the biases are related to overconfidence in the technological process.

It should also be noted that careful standardization for the definition of uncertainties and the direction of their scale must be used. One possibility to make the interpretation more consistent is to always use a *positive* form: in this case *stagnation of phenomenon x* is not a suitable form for an uncertainty, but *progress of phenomenon x* is. Only in this way the observed mean outcome offers any usable information about systematic errors.

(2) *Variances*. The uncertainty outcome data can also be used to measure the variance of uncertainty outcomes. This sample variance provides information about our assumptions about the probability for extreme events. If the sample variance exceeds one, we systematically underestimate the probability for extreme outcomes. Conversely, when the sample variance is below one, we tend to overestimate the probability for extreme outcomes. The correct standard deviation is required in the analysis so that the range of each scenario set becomes relevant, given the desired weights for scenario diversity and plausibility. It should again be noted that the test sample should only include uncertainties for which independence can be reasonably assumed. In this way the standard chi-squared statistic can be used to test whether the assumed variance holds in the data. As with t-tests, the uncertainties can also be divided into separate sets and test the variances of the sets separately. Assumptions should again be iterated based on the results.

(3) *Correlations*. We also want information about our ability to identify co-dependencies. In our case we have only one observation for each multivariate distribution, which implies the sample correlation (calculated based on the zero mean) between each uncertainty pair is either 1 or −1. We then compare the observed sign to the sign of the approximated correlation and infer whether a *sign reversal* happened between the variables, i.e. whether the observed sign did not match the approximated sign. We then count the total number of sign reversals in the data. Ideally, the relative number of sign reversals should be below 50%. This would imply that on average, we are able to approximate the sign of co-dependencies correctly. Although the observed sign reversals are not always independent, the relative number might give some directional knowledge about the ability to predict co-dependencies. However, by utilizing the full correlation structure, it is also possible to form a relatively independent set of variable pairs so that the number becomes more reliable.

The correlations can also be rounded to the nearest decimal, and these correlation levels can be tested separately. The expected number of sign reversals is known for each correlation level, and the observed number is then compared to the expected number to decide whether our correlation assessments on that level systematically under- or overestimate the true correlation.

Ex posteriori testing of parameter assumptions does not require phrasing any additional questions for experts, as there are no new numerical expert assessments to be done. The analyst(s) performing the calculations should only gather the earlier assessments and exclude correlated uncertainties from the test data. Lastly, the test results should be presented to the experts in an understandable form to effectively eliminate possible systematic biases.

4.3. Forecasting skill

4.3.1. Definition of forecasting skill

Lastly, we can test for the overall forecasting skill of our model and our assumptions. Forecasting skill refers to the ability of a forecaster or a forecasting method to accurately predict future events or values. This skill is measured by scoring rules that are mathematical functions used to evaluate and compare the performance of different forecasting models by quantifying the accuracy of their predictions (Gneiting & Raftery, 2007; Wheatcroft, 2019). The rules assign scores based on the alignment between predicted probabilities and actual outcomes, providing a systematic way to assess and improve forecasting accuracy. On the other hand, if the skill scores are systematically poor, the method itself might not function properly in the given context.

Skill scores usually quantify the improvement in performance when using one forecasting system compared to another, expressed as a fraction of the improvement attained by perfectly predicting the outcome. In our case the outcome is assumed to be random, and perfect forecasting skill is ruled out. Consequently, the value of comparing the accuracy of our distributional assumptions to perfect prediction is questionable. As Wheatcroft (2019) argues, the skill score form of scoring rules also destroys the useful interpretation in terms of the relative skill levels of two forecasting systems, and the skill score forms of several established scoring rules are biased in small samples. Thus we will only measure the *relative* skill levels of two systems.

We will compare the forecasting ability of our approximated distributions to *random guessing*, which in mathematical terms in interpreted as an uniform distribution over the N-dimensional hypercube. Our relative skill score then measures the gain in forecasting ability relative to pure randomization. This provides an intuitive interpretation for the score. When the relative skill score is high, the expert assessments are accurate and the multinormal approximation provides valuable information. Conversely, when the skill score is low, either the expert

assessments are poor or the utility of the approximation framework presented in this paper is questionable.

A scoring rule is deemed *proper* when its optimal outcome, in terms of expected performance, aligns with a perfect probabilistic forecast (Gneiting & Raftery, 2007). In other words, proper scoring rules are designed to be maximized when the forecast accurately reflects the true distribution from which the outcome is derived. We utilize two different scoring rules that are known to be proper: the *Brier score* and the *logarithmic score* (also known as the *ignorance score*).

4.3.2. Brier score

The Brier score, developed by Glenn W. Brier in 1950 (Brier, 1950), has emerged as a prominent metric for quantifying the accuracy of probabilistic forecasts, and it has been popularized especially by Tetlock (2005), Tetlock and Gardner (2015). The score is designed for discrete cases and it measures the mean squared difference between predicted probabilities and the actual outcomes.

To be able to use the Brier score we have to divide the continuous variable space into distinct cells (or *neighborhoods*) that each form a relatively similar subset of possible states of the future. Each cell has its own probability which is based on our distributional assumptions. A natural way to define the neighborhoods is to use so called Voronoi cells. This means that given some set of scenarios $S = \{\mathbf{u}_1, \dots, \mathbf{u}_M\}$, Voronoi cell $j \in \{1, \dots, M\}$ corresponding to scenario $\mathbf{u}_j$ consists of every point for which $\mathbf{u}_j$ is the nearest scenario from the given set. In other words, in Voronoi cell j the distance to $\mathbf{u}_j$ is less than or equal to the minimum distance to any other site $\mathbf{u}_k$, where $j \neq k$. Thus, Voronoi cell j, denoted by C_j, is formally defined as

$$C_j = \{\mathbf{u} \in \mathbb{R}^N \mid D_E(\mathbf{u}, \mathbf{u}_j) \leq D_E(\mathbf{u}, \mathbf{u}_k) \text{ for all } j \neq k\}. \tag{15}$$

We use the lower index E and define that the set of data points $S_E = \{\mathbf{u}_1, \dots, \mathbf{u}_M\}$ that corresponds to the Voronoi cells is called the *evaluation set* to make a distinction between the data points used in evaluation and the optimal scenario set which is used in the actual scenario analysis. Here, M does not necessarily equal the number given by Eqs. (9) and (10).

Given the definition of Voronoi cells, each random outcome lies in one cell. The theoretical probability for each cell $j \in \{1, \dots, M\}$, denoted by $p_k = \mathrm{P}(\mathbf{U} \in C_k)$, is then defined as the cumulative distribution function over cell j, i.e.

$$\mathrm{P}(\mathbf{U} \in C_j) = \int_{C_j} f(\mathbf{u})\, d\mathbf{u}, \tag{16}$$

where f is the probability density function of the assumed distribution. Now each cell can be given a probability, and the aggregate probability of all cells is 100 percent.[5] However, working with theoretical distributions is a complex task, and thus probabilities for each cell can also be calculated using Monte Carlo simulations. By generating a large number of outcomes from the assumed distribution, relative frequencies for each cell give an approximation.

How then should the evaluation set S_E be selected? There are several possibilities, but we propose that the points are selected by dividing the N-dimensional space into a uniform grid. This involves creating a structured arrangement of equally sized and shaped cells or regions that cover the entire space, and every dimension (or axis) is divided into a consistent number of subdivisions. Although the grid can be made as granular as desired, we propose that each axis is divided into two parts: the positive part and negative part. Thus the grid regions are equal to the 2^N orthants in the variable space. The evaluation set that corresponds to the orthants is obtained by:

$$S_E = \{(u_1, \dots, u_N) \mid u_i \in \{-a, a\} \text{ for } i = 1, \dots, N\}, \tag{17}$$

where $0 < a \leq 2$.

In the following, we will use the lower index k to denote the forecasting instance, i.e. the topic. As presented in e.g. Wheatcroft (2019), the formula for the Brier score in forecasting instance k is given by

$$BS_k = \sum_{j=1}^{M_k} (p_{kj} - o_{kj})^2, \tag{18}$$

where $o_{kj} = 1$ when the actual outcome lies in cell j, and $o_{kj} = 0$ otherwise. When the orthants are used as cells, $M_k = 2^{N_k}$. The smaller the score the better the forecasts are.

When the outcome is distributed uniformly, all orthants are equally probable. Thus the reference score corresponding to random guessing is given by

$$BS_k^r = \sum_{j=1}^{M_k} (1/M_k - o_{kj})^2, \tag{19}$$

The Brier score is usually averaged over several forecasting instances. Because the number of cells varies between instances, we cannot simply calculate the average of all Brier scores without losing information. Thus we first calculate the relative skill score in each instance separately and then calculate the average. The relative skill score in instance k is given by $(BS_k^r - BS_k)/BS_k^r$, and thus the final skill score over all instances is given by

$$BS = \frac{1}{Q} \sum_{k=1}^{Q} \frac{BS_k^r - BS_k}{BS_k^r} = 1 - \frac{1}{Q} \sum_{k=1}^{Q} \frac{BS_k}{BS_k^r} \tag{20}$$

When the approximated multinormal distribution outperforms the uniform distribution, the value of BS is greater than 0. The score indicates the relative gain in Brier scores as a fraction of the score obtained from random guessing. When then value is 1, we have perfect ability to forecast. However, as explained above, this is not even theoretically achievable in the framework presented.

4.3.3. Logarithmic score

The logarithmic score or ignorance score is another proper scoring rule. Contrary to Brier scores it can be used for both discrete and continuous cases. Here we present its use in a continuous setting so that we do not have to make any discretization in our variable space. The logarithmic score is calculated as the logarithm of the probability estimate for the actual outcome, and in the continuous case this means taking the logarithm of the probability density function at the outcome. Mathematically, the score at forecasting instance k is given by

$$LS_k = -\ln(f_k(\mathbf{u}_k^A)), \tag{21}$$

where f_k is the probability density function for the approximated multinormal distribution and $\mathbf{u}_k^A$ is the actual outcome. This score tends to penalize errors on low-probability events more heavily than high-probability events. Again the lower the score the better the forecasts are.

The reference forecast system is again defined as random guessing, i.e. the reference distribution is the uniform distribution over the relevant part of the variable space. The reference score is then given by

$$LS_k^r = -\ln(1/V_k), \tag{22}$$

where $1/V_k$ is the probability density function for the uniform distribution and V_k is the volume of the relevant part of the variable space. Although the actual outcome for each uncertainty can also lie outside

[5] It should be noted that although it might seem tempting, we recommend that discretization is not used to assess discrete scenario probabilities. This is due to the fact that given a scenario data point that defines a Voronoi cell, the assigned probability depends both on the plausibility of the scenario and the granularity of cell division. Thus the probability assigned to a scenario point does not provide any value if the full cell division is not simultaneously presented.

Table 1
Assessment of thresholds.

	Minimal	Expected	Maximal
Development of EV Substitutes	Growth of investments in EV substitutes such as fuel cells and synthetic fuels have stopped and started to decline. No commercial applications for consumers.	Steady growth rate in both investments and the number of commercialized products. Globally, there are 13 million hydrogen fuel cell vehicles.	The market growth rate has increased. The market size has doubled each year. Some commercial products available and their price is competitive relative to EVs.
Government Incentives for EVs	The timeline for banning ICE cars has been extended in the EU and several other areas. Almost no new financial incentives to promote EVs have been implemented.	The timeline for banning ICE cars is in place. Some EV subsidies and tax schemes have been phased out with a shift towards budget-neutral "feebate" programs and stringent vehicle efficiency/CO2 standards.	Almost all developed areas in the world have announced timelines for banning internal combustion engine (ICE) vehicles. The subsidies for EVs have doubled.
Development of Batteries	Almost no development in performance, range, and sustainability of batteries. New alternative battery technologies have failed.	Steady development and some emerging new technologies.	The price of batteries has fallen by a third. Energy density has doubled. There are some commercially available alternative battery technologies.
Price of Electricity	The price of electricity has fallen sharply, 50% on average.	No long-term changes or trends in electricity prices.	The price of electricity has doubled on average.
Development of Ride-sharing and Public Transport	Ride-sharing apps have been banned in several new countries. No major new public transport projects.	Steady development of public transport in major cities but the number of passengers and trips has remained fairly constant. There are some new ride-sharing apps and their popularity has grown steadily.	Number of trips (public transport) has grown around 20% in major cities. Record number of ongoing investments in new public transport projects. Ride-sharing market has grown 50% per year.
EVs' Appeal to Consumers	The price of electric cars has not decreased. Attitudes towards electric cars and their owners are mainly negative.	Weak negative development in prices. Moderate positive trends in attitudes and overall appeal.	Attitudes towards EVs have become extremely positive. ICEs are seen as unresponsible. There is huge consumer demand for Evs regardless of their price and performance.
Availability of Rare-Earth Metals	Availability has deteriorated. Average prices have doubled.	Weak upward price trends in rare-earth metals.	Weak negative trends in rare-earth metal prices. Availability is no more a concern.
Concerns About Climate Change	The visibility of Climate change in news, social media etc has dramatically fallen.	Climate change is as big a concern as previously.	The concerns are widespread and visible everywhere. Climate anxiety is widespread among teens. The number of climate change-related news has doubled.

the interval $[-2, 2]$, we use the volume of the N-dimensional hypercube formed by this interval. Thus we have that $V_k = 4^{N_i}$. This selection is somewhat arbitrary: the larger the selected hypercube, the smaller reference scores we get and the more inflated relative skill scores we will obtain. However, given the definition of the endpoints for the interval, the selection $[-2, 2]$ is the most intuitive.

The relative skill score in each forecasting instance k is again given by $(LS_k^r - LS_k)/LS_k^r$, and the final skill score over all instances is given by

$$LS = \frac{1}{Q}\sum_{k=1}^{Q}\frac{LS_k^r - LS_k}{LS_k^r} = 1 - \frac{1}{Q}\sum_{k=1}^{Q}\frac{LS_k}{LS_k^r}. \tag{23}$$

As with Brier scores, when the approximated multinormal distribution outperforms the uniform distribution, the value of LS is greater than 0. The score indicates the relative gain in logarithmic scores as a fraction of the score obtained from random guessing.

5. Case example: Electric cars 2030

The following section demonstrates the usage of the presented method in scenario production with a case example. The scenarios explore possible futures regarding the global market for electric vehicles (EVs) in 2030 on a very general level. The results have been visualized using a "spiderweb" image, and we provide a brief narrative as an interpretation for each scenario. The goodness of the scenario set is based on the previously explained mathematical optimality. The set was originally formed by a team of futurists at Futures Platform in 2024.

5.1. Workflow

The workflow was built to match the internal scenario production process at Futures Platform. A detailed analysis of the practical use of the method in a workshop setting (e.g. with several different stakeholders) is left for future research. However, the general steps presented below can be directly applied to several other contexts and settings:

- Selection of the topic, time frame, and doing the basic research to understand how the main trend is evolving, what are the strongest drivers, and what kind of things are causing uncertainty to the issue.
- Identification and prioritizing the most relevant or impactful uncertainties.
- Assessment and documentation of the interpretations for the thresholds (minimal, expected, and maximal levels). Finding quantitative and qualitative data to support the assessments.
- Assessment and documentation of correlations between uncertainties. Finding quantitative data whenever possible.
- Computation of eigenvalues, selection of the number of scenarios and computation of the resulting scenarios with a desired level of diversity.
- Interpretation and visualization of the numerical results.
- Assessment and compilation of impacts of the states of the uncertainties into a coherent storyline. Building the final scenario narratives and development paths.

5.2. Identification of uncertainties

A set of initial uncertainties were first identified by a single (main) analyst who performed a PESTLE-based horizon scan. These initial uncertainties were then passed to an expert panel for an internal review and discussion which augmented, combined and pruned the original set of uncertainties. As a result, the following final set of uncertainties was formulated:

- Development of EV Substitutes (U1)

Table 2
Uncertainty correlation table.

	U1	U2	U3	U4	U5	U6	U7	U8
U1	1	−0.3	−0.4	0.3	0.1	−0.3	−0.4	0.2
U2	−0.3	1	0.3	−0.1	−0.1	0.2	0.2	0.3
U3	−0.4	0.3	1	0.3	0	0.3	−0.4	0.2
U4	0.3	−0.1	0.3	1	0.2	−0.5	−0.2	0.1
U5	0.1	−0.1	0	0.2	1	−0.1	−0.3	0.3
U6	−0.3	0.2	0.3	−0.5	−0.1	1	0.1	0.4
U7	−0.4	0.2	−0.4	−0.2	−0.3	0.1	1	−0.1
U8	0.2	0.3	0.2	0.1	0.3	0.4	−0.1	1

- Government Incentives for EVs (U2)
- Development of Batteries (U3)
- Price of Electricity (U4)
- Development of Ride-Sharing and Public Transport (U5)
- EVs' Appeal to Consumers (U6)
- Availability of Rare-Earth Metals (U7)
- Concerns About Climate Change (U8)

5.3. Formulation of thresholds

Next, the interpretations for the thresholds −2 (minimal), 0 (expected) and 2 (maximal) were documented. These assessments were formed by the main analyst who gathered forecasts, analyses and other information which supported the assessment of expected outcomes. For example, the expected size for the market of EV substitutes was assessed based on a set of external market forecasts that were available. The results are presented below in Table 1.

5.4. Correlation analysis

As with uncertainty identification, the correlations between uncertainties were first assessed by the main analyst. A suitable accuracy was selected to be one decimal. The main questions that guided the analyst through the process were: (1) Does uncertainty A have a positive or negative effect on uncertainty B? (2) Does uncertainty B have a positive or negative effect on uncertainty A? (3) Is there a third factor that influences both uncertainties? (4) How strong are these possible effects?

Most of the initial correlation assessments were based on expert judgement and some were directly supported by quantitative data. Supportive qualitative information was also sought in cases with no numerical data. For example, time series were used that supported the idea of a positive correlation between prices of electricity and rare-earth metals. Consequently, the correlation between the price of electricity and the availability of rare-earth metals was assessed to be negative.

The correlation table was then passed on to an internal review process, where other analysts were able to ask for clarification and suggest changes based on their own thoughts and observations. Thus, the process again included active discussion and the numerical assessments became iterative. A Delphi process could also have been used, but in this case numerical values were iterated only through discussion. The final correlations are presented in Table 2.

5.5. Numeric scenario results

Next, the eigenvalues of the correlation matrix were computed to decide the final number of scenarios. In this case, the value $c = 2/3$ was selected by the main analyst, representing the midpoint of the possible range of the parameter. This implies that all principal components with variances less than two-thirds of the variance of the first component were discarded. This resulted in using the two first principal components, implying four scenarios. The scree plot also supported the selection of two principal components. On the other hand, using the Kaiser criterion or e.g. $c = 0.5$ would have resulted in six scenarios.

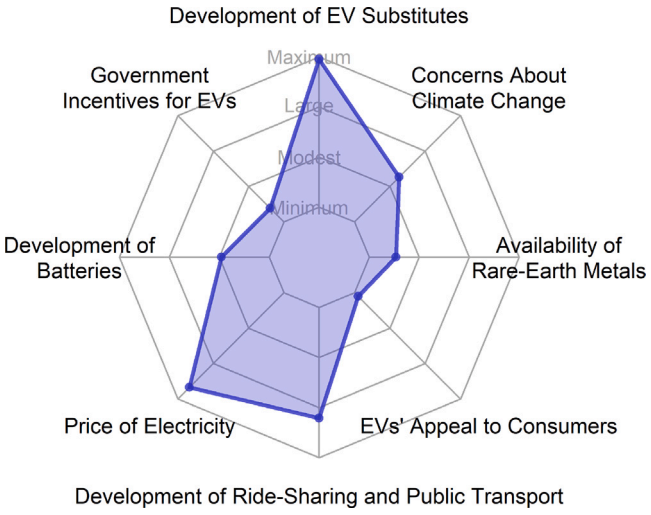

Fig. 3. Spiderweb visualization of Scenario 1.

A large weight was put on scenario diversity, as the team wanted scenarios that cover the range for possible outcomes maximally. Thus $\alpha_d = 2.6$ was selected, as this resulted in uncertainty values close to the minimal and maximal values in each scenario.

The resulting scenario vectors' numeric values are rounded and presented in Table 3, which is a heat map. The color coding is the following: $[-2, -1)$: blue; $[-1, 0)$: light blue; $[0, 1)$: light red; $[1, 2]$: red. These correspond to minimal, modest, large and maximal levels of development.

5.6. Scenario narratives with visualizations

Finally, the results in Table 3 were briefly interpreted and visualized with a simple radar chart (or "spiderweb") which presents all uncertainty values on a scale from minimum to maximum in a polygon. Other visualization possibilities could have included e.g. more precise heat maps, bar patterns or parallel coordinates plots.

The resulting scenario narratives were first formed by the main analyst, and this was again followed by an internal review process.

5.6.1. Scenario 1: Mineral shortages and the rise of synthetic fuels hamper the adoption of electric vehicles

The scarcity of raw materials, including lithium, cobalt, and rare earth metals, intensifies, putting heavy upward pressure on EV prices and strongly hurting their appeal. The increased demand for electricity due to global electrification leads to higher costs associated with electricity generation and distribution. High electricity prices and scarcity of materials lead to low investments in battery development and instead increased investments in alternative transportation technologies. Technological breakthroughs in synthetic fuel production gain momentum. Innovations in electrolysis, gasification, and Fischer–Tropsch synthesis make the process more efficient and cost-effective. These advancements enable the conversion of renewable energy, such as solar

Table 3
Uncertainty values.

	u_1	u_2	u_3	u_4
Development of EV Substitutes	1.94	−1.94	−0.11	0.11
Government Incentives for EVs	−1.49	1.49	−0.85	0.85
Development of Batteries	−0.72	0.72	−1.94	1.94
Price of Electricity	1.56	−1.56	−0.96	0.96
Development of Ride-Sharing and Public Transport	0.95	−0.95	−1.17	1.17
EVs' Appeal to Consumers	−1.87	1.87	−0.79	0.79
Availability of Rare-earth Metals	−1.29	1.29	1.48	−1.48
Concerns About Climate Change	−0.32	0.32	−1.81	1.81

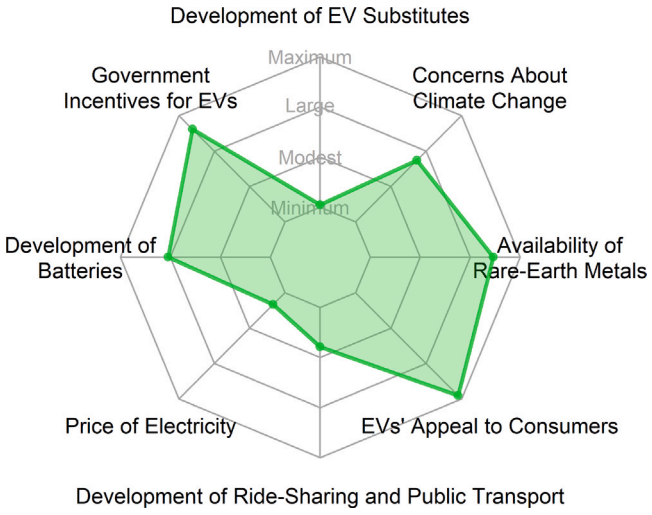

Fig. 4. Spiderweb visualization of Scenario 2.

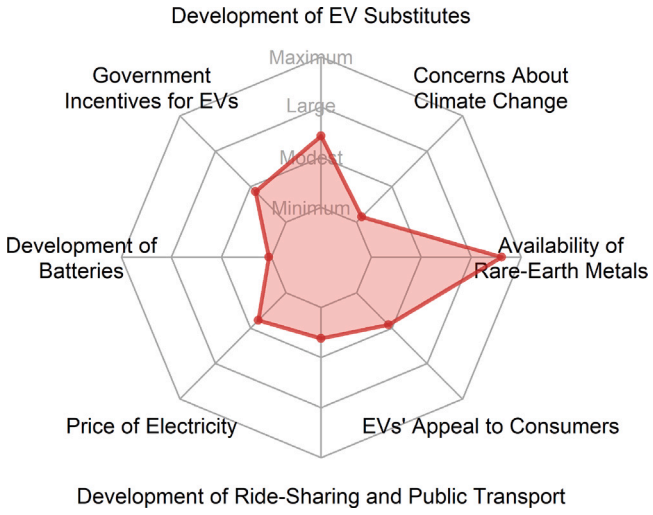

Fig. 5. Spiderweb visualization of Scenario 3.

and wind power, into synthetic fuels with higher energy densities. The availability of cost-effective synthetic fuels offers an alternative to EVs, addressing concerns over high and volatile electricity prices. Ride-sharing and public transport gain popularity as affordable and eco-friendly alternatives to private car ownership, leading governments and municipalities to invest in expanding transportation networks, improving infrastructure, and implementing innovative ride-sharing services to meet growing demand. Financial incentives for EVs are mostly halted (see Fig. 3).

5.6.2. Scenario 2: Total electrification of traffic and society

Widescale electrification of society gains momentum. Investments pour into mining of critical minerals needed for renewable energy and EVs as part of the great power competition for the future of green energy. The combination of low electricity prices and comprehensive government incentives and policies creates a compelling proposition for electric vehicle ownership. The escalating interest in electric vehicles amplifies the demand for battery minerals, reshaping global commodities markets and incentivizes new mining projects. Multiple breakthroughs in battery technology occur. In addition, advancements in materials, recycling methods and improved material usage lead to more efficient and sustainable batteries, further enhancing the appeal of EVs. Governments subside low-interest loans and induce tax-breaks to encourage widespread EV adoption, resulting in a surge in EV purchases. The development of EV substitutes stagnates and no major public transport projects are implemented (see Fig. 4).

5.6.3. Scenario 3: Electric vehicles cannot counter the attractiveness of internal combustion engines

The electric car market faces a downturn as various factors converge to create a challenging landscape for the once-promising industry. Governments continue to push for a green transition and offer some modest incentives for EV adoption. Despite these efforts, the incentives struggle

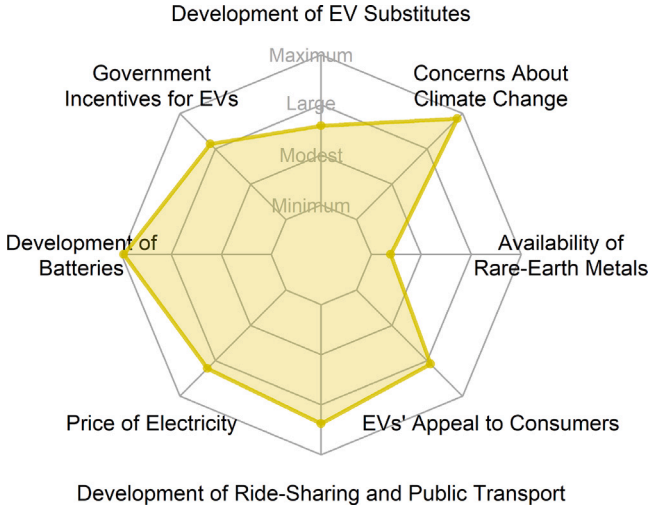

Fig. 6. Spiderweb visualization of Scenario 4.

to counter the popularity of traditional internal combustion engines. Concerns about climate change have begun to fade from the public consciousness, and as immediate environmental threats and resource scarcity seem to recede, consumers become less inclined to prioritize eco-friendly transportation options. The development of ride-sharing, public transport and EV substitutes is sluggish, advancements in battery technology lag behind expectations, and the limited driving range of electric cars and the time-consuming charging process further deter potential buyers. Consequently, the general public finds traditional combustion engine vehicles to still be more attractive (see Fig. 5).

5.6.4. Scenario 4: Climate awareness reshapes the transportation sector

Heightened concerns about climate change capture the public's attention. Extreme weather events, scarcity of raw materials and environmental degradation fuel a collective sense of urgency to combat climate change, making consumers more environmentally conscious. This shift in mindset significantly increases the appeal of electric vehicles, as people seek greener alternatives to traditional combustion engine vehicles. Simultaneously, high electricity prices create a hurdle for the industry, and governments worldwide implement robust incentives to promote electric vehicle adoption. The scarcity of minerals leads to increased competition for these resources, driving up production costs for electric cars. However, it also prompts major resource-saving innovation in battery technology, encouraging the development of alternative materials and more sustainable practices. The largest challenge for the EV market is the development and rising popularity of ride-sharing and public transport, fueled by environmental concerns. Combined with high electricity prices, this lowers the overall demand for private vehicles (see Fig. 6).

5.7. Testing the assumptions

With one single approximated distribution it is impossible to calculate e.g. forecasting skill, as this requires more approximated distributions than just one. However, Futures Platform has began collecting test data to evaluate its future-related assumptions. This data consists of 30 different distributions with a time frame of 2 years into the future and with an extremely diverse range of future-related phenomena. The distribution for electric vehicles is one of the selected topics. The full list of topics covered, their chosen uncertainties, numerical assessments, outcomes and the performance of the framework are presented in future research. The outcomes will be analyzed in 2026.

The scenarios built in this section correspond to the year 2030. Thus the exact thresholds (−2, 0 and 2) of the selected uncertainties differ in the test data compared to Table 1, as the test is built with a time frame of only two years. However, Table 1 can be used to demonstrate how the actual outcomes would be documented at the chosen time period: was the development of each uncertainty stronger or weaker than the expectation? How close was the outcome to the expectation or the minimum or maximum wild card outcome? For example, if the price of electricity would have almost doubled, the outcome value of this uncertainty could be set to something between 1.5 and 2 based on the assessment in the table.

6. Discussion

In this article, we have presented a new practical framework for decomposed and exploratory scenario analysis that utilizes simple statistical concepts: the normality of possible outcomes and the linearity of co-dependencies. In this way scenario building is brought closer to concepts common among other more established scientific fields, and we bring some elements from the forecasting tradition into exploratory scenario analysis. As Roponen and Salo (2023) argue, qualitative and quantitative scenario methods are not fundamentally at odds, but rather complementary. We have also argued that as scenario building often relies on purely subjective assessments, more emphasis should be put on ex posteriori evaluation of our assumptions and their performance. For this purpose we have proposed some simple tests and metrics suitable in the presented framework. We have also presented the usage of the method by building four simple scenarios on the future of electric cars.

The method we have presented maximizes the plausibility, measured by Mahalanobis distances, and the diversity, measured by distances, of the scenario set generated. Therefore the scenarios are not only plausible but they cover a maximally wide range of future possibilities so that the scenarios are not just variations on similar themes. This also makes choosing a set of scenarios from the large number of possible candidates efficient and consistent. Moreover, the framework has many advantages relative to earlier categorical methods with the same objectives. Firstly, our method is analytically and computationally relatively easy to conduct due to a smaller number of co-dependency assessments. This enables the analysts to work with all the critical uncertainties without the need to censor or combine uncertainties due to lack of time, computing power or other constraints. Secondly, the results are much more easily visualizable compared to earlier methods. Thirdly, the statistical parameters we utilize are measures which in some cases can be directly estimated or supported using available data, and they have real statistical interpretations, contrary to different arbitrary scales of consistency. This also allows common statistical tests to be used on the results to evaluate their performance. However, the availability of numerical data largely depends on the topic. We assess that in most scenario settings direct estimates cannot be utilized, but there are also opposite examples, such as scenarios with a large number of economic and financial variables.

An additional strength of the framework is its generalizability. If the scale of the Gaussian variables is divided into separate intervals, the method in effect becomes the same as other decomposition methods with discrete uncertainty outcome levels. If the number of uncertainties is 2, the method is reduced to the so called *deductive*, or the 2×2, scenario framework.[6] which is widely used among different fields (see e.g. Ramirez et al., 2015; van der Heijden, 2005). The method can be used for purely numerical applications so that the parameters are estimated from data, but also for purely qualitative cases and all cases in between.

Naturally, a clear shortcoming in our approach is that biases in expert judgement are tackled by a method that draws mainly on expert judgement. However, relative to other methods, our approach can be rationalized also by using the famous bias–variance trade-off as an analogue. This trade-off is a fundamental concept that describes the balance between two sources of error affecting model performance: high bias occurs when a model is too simple, leading to underfitting, and high variance happens when a model is too complex, causing it to overfit and perform poorly on new data. Achieving the right balance between bias and variance is crucial for developing a model that generalizes well to new data. With regard to scenario analysis, the approach must not have too few numerical assessments (underfitting), and not too many (overfitting). We believe that with categorical variables there is a clear risk of overfitting, as there is a huge number of required expert assessments which lead to complex, nonlinear dependencies between the variables. We also believe that our approach achieves the right balance, because co-dependencies are assumed to be linear, as is the case in standard regression analysis. Consequently, scarce future knowledge is condensed into a few simple statistical parameters. These parameters correspond to factors about which there might be important directional knowledge: (1) what is most probably going to happen, (2) how different factors are dependent on each other and (3) what extreme possibilities there are regarding the factors.

Although the analysis process and the selection of uncertainty level combinations is more transparent than in many other decomposed methods, there is still room for more. Assessment of correlations and the interpretation of the uncertainty scales $[-2, 2]$ require a robust, reliable and transparent analysis process, especially when the uncertainties cannot be easily expressed in a numerical form. The same applies to the building of scenarios based on the resulting numerical uncertainty values: how should the numerical results be translated into real language so that numbers and qualitative language match each other? The implications of the actual outcome, expressed by the numerical vector, are something the analyst still has to interpret based on pure expert knowledge.

[6] This can be noticed also in Figs. 1 and 2: we have selected four scenarios with all four different combinations of binary uncertainty levels: low–low, low–high, high–low and high–high.

There are also other paths for future research that remain to be further analyzed, such as the deployment of the method in a workshop setting. In a workshop, the participation of different stakeholders, active discussion and Delphi-based estimates might provide additional robustness for the numerical assessments. The main challenge, in turn, might be related to understandably defining correlations and phrasing correct questions that lead the participants through the analysis steps.

Moreover, in this article we have restricted the analysis to explorative project goals where the uncertainties are external, i.e. outside the scope of influence of the actor. However, there are no specific reasons why internal factors, i.e. factors that are controllable by the actor, should be left out. If there exists co-dependencies between the level of internal and external factors, correlations can be assessed similarly as with other variable pairs. In this case, however, the causation should be assumed to run from the internal factor to the external uncertainty, as the internal factors are not assumed to be random. The analysis could then be performed by fixing the level of the internal factors into some desired level, and then proceeding similarly with the same goals of diversity and plausibility.

This second type of explorative scenarios is sometimes called *strategic* (Börjeson et al., 2006). These scenarios focus on internal factors, while taking external aspects into account and respond to the question: *What can happen if we act in a certain way?* The aim of explorative strategic scenarios is to describe a range of possible consequences of strategic decisions regarding the issue at stake. In other words, they study and describe how the consequences of a decision or policy can vary depending on which future development unfolds (Börjeson et al., 2006).

The external variables can naturally also include strategic choices of other actors such as competitors, which opens very intriguing possibilities. By including anticipated competitor behavior or other analogous external factors, the analysis approaches game theory or normative project goals that respond to the question: *how can a specific target be reached?* (Börjeson et al., 2006). However, this not only applies to game theory but also to its Russian counterpart, reflexive control which seeks to influence the decision making of the opponent (Vasara, 2020), meaning that the proposed method not only has the potential to explore the future, but through it to explore the ways to anticipate and direct the present actions of others. Altogether, the generalizability of the framework makes it highly interesting and applicable for different disciplines such as foresight, business strategy, finance, strategic intelligence analysis, and international relations.

CRediT authorship contribution statement

Eljas Aalto: Writing – original draft, Methodology. **Tuomo Kuosa:** Writing – review & editing, Supervision. **Max Stucki:** Writing – review & editing, Validation.

Acknowledgments

This research has been funded by The University of Turku Graduate School (UTUGS) and Futures Platform Ltd.

Appendix. Proof of Result 1

We assume the optimization is carried out iteratively, implying that scenarios are added to the set one by one by optimizing each selection separately. When the number of scenarios is 2, it is easy to see that given some fixed value for G_p, the two scenarios must lie on opposite ends of the first principal component, i.e. the direction along which the distribution has the largest variance. This will ensure that given some level of plausibility, the distance between the two scenarios is maximal. It is also easy to notice that the Mahalanobis distance for both scenarios must be equal in the optimum: otherwise it would be possible to increase G_d without having an effect on G_p. Thus we have that $\mathbf{u}_2 = -\mathbf{u}_1$.

Now suppose that $\lambda_2/\lambda_1 \geq 1/3$, where λ_1 and λ_2 are the first and second eigenvalues of the correlation matrix $\mathbf{\Sigma}$, i.e. the variances of the first and second principal component. When the number of scenarios is 4 and $\mathbf{u}_1$ and $\mathbf{u}_2$ lie on the first principal component as presented, the two additional scenarios are similarly placed on the opposite ends of the second principal component. Given some fixed value for G_p, if the third and fourth scenario did not lie along the second principal component, G_d could be increased without having an effect on G_p. For the same reason, the Mahalanobis distance must again be equal for all scenarios, and thus $\mathbf{u}_3 = -\mathbf{u}_4$. The condition $\lambda_2/\lambda_1 \geq 1/3$ ensures that the minimum pairwise distance is given by the distance between $\mathbf{u}_1$ and $\mathbf{u}_3$ (or the three other pairs with equal distance), and not between $\mathbf{u}_3$ and $\mathbf{u}_4$.

Similarly, when the number of scenarios is 6 and $\lambda_3/\lambda_1 \geq 1/3$, the same logic applies. In this case $\mathbf{u}_5 = -\mathbf{u}_6$ and the minimum pairwise distance is given e.g. by the distance between $\mathbf{u}_3$ and $\mathbf{u}_5$.

Next, we will transform the scenarios into the principal component space. Because the optimal scenarios must lie on the principal components, the transformed vector will have a single non-zero component, which corresponds to the principal component on which the scenario lies. We also know that if scenario $\mathbf{u}_j$ lies on principal component i, then $D^2_M(\mathbf{u}_j;\mathbf{\Sigma}) = z^2_{ji}/\lambda_i$. Next, we will demonstrate the solution for each M separately

When $M = 2$:

The objective function P is in the optimum given by

$$P(S;\mathbf{\Sigma}) = 2\alpha_d|z_{11}| - \alpha_p\frac{z^2_{11}}{\lambda_1}. \tag{A.1}$$

Assuming that $z_{11} > 0$, taking the derivative with respect to z_{11} and setting the derivative to zero leads to the following solution:

$$z_{11} = \frac{\alpha_d}{\alpha_p}\lambda_1. \tag{A.2}$$

Lastly, we use $z_{21} = -z_{11}$.

When $M = 4$:

We have that in the optimum $D^2_M(\mathbf{u}_1;\mathbf{\Sigma}) = D^2_M(\mathbf{u}_3;\mathbf{\Sigma})$ which implies that

$$z^2_{32} = \frac{\lambda_2}{\lambda_1}z^2_{11}. \tag{A.3}$$

The objective function is in the optimum given by

$$P(S;\mathbf{\Sigma}) = \alpha_d\sqrt{z^2_{11}+z^2_{32}} - \alpha_p\frac{z^2_{11}}{\lambda_1}. \tag{A.4}$$

Substituting Eq. (A.3) into (A.4), then taking the derivative with respect to z_{11}, setting it to zero and assuming $z_{32} > 0$ leads to the following solution:

$$z_{11} = \frac{\alpha_d}{\alpha_p}\frac{\sqrt{\lambda_1(\lambda_1+\lambda_2)}}{2} \tag{A.5}$$

$$z_{32} = \frac{\alpha_d}{\alpha_p}\frac{\sqrt{\lambda_2(\lambda_1+\lambda_2)}}{2}. \tag{A.6}$$

Lastly, we again use $z_{21} = -z_{11}$ and $z_{42} = -z_{32}$.

When $M = 6$:

We have that in the optimum $D^2_M(\mathbf{u}_1;\mathbf{\Sigma}) = D^2_M(\mathbf{u}_3;\mathbf{\Sigma}) = D^2_M(\mathbf{u}_5;\mathbf{\Sigma})$ which implies that Eq. (A.3) again holds and in addition,

$$z^2_{53} = \frac{\lambda_3}{\lambda_2}z^2_{32}. \tag{A.7}$$

The objective function is in the optimum given by

$$P(S;\mathbf{\Sigma}) = \alpha_d\sqrt{z^2_{32}+z^2_{53}} - \alpha_p\frac{z^2_{53}}{\lambda_3}. \tag{A.8}$$

Substituting Eq. (A.7) into (A.8), then taking the derivative with respect to z_{32}, setting it to zero and assuming $z_{53} > 0$ leads to the following solution:

$$z_{11} = \frac{\alpha_d}{\alpha_p} \frac{\sqrt{\lambda_1(\lambda_2 + \lambda_3)}}{2} \tag{A.9}$$

$$z_{32} = \frac{\alpha_d}{\alpha_p} \frac{\sqrt{\lambda_2(\lambda_2 + \lambda_3)}}{2} \tag{A.10}$$

$$z_{53} = \frac{\alpha_d}{\alpha_p} \frac{\sqrt{\lambda_3(\lambda_2 + \lambda_3)}}{2} . \tag{A.11}$$

Lastly, we again use $z_{21} = -z_{11}$, $z_{42} = -z_{32}$ and $z_{63} = -z_{53}$.

Now by setting $\alpha_p = 1$ we obtain the results given by Eqs. (11)–(13).

References

Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. *Quarterly Journal of Economics*, *120*(2), 701–728.

Aguilar, F. (1967). *Scanning the business environment*. New York: Macmillan.

Ahrenshop, M., Golden, M., Gulzar, S., & Sonnet, L. (2023). Inaccurate forecasting of a randomized controlled trial. *Journal of Experimental Political Science*.

Barrios, M., Guilera, G., Nuño, L., & Gómez-Benito, J. (2021). Consensus in the delphi method: What makes a decision change? *Technological Forecasting and Social Change*, *163*, Article 120484.

Börjeson, L., Höjer, M., Dreborg, K.-H., Ekvall, T., & Finnveden, G. (2006). Scenario types and techniques: Towards a user's guide. *Futures*, *38*(7), 723–739.

Brauers, J., & Weber, M. (1988). A new method of scenario analysis for strategic planning. *Journal of Forecasting*, *7*(1), 31–47.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Bunn, D. W., & Salo, A. A. (1993). Forecasting with scenarios. *European Journal of Operational Research*, *68*(3), 291–303.

Carlsen, H., Eriksson, E. A., Dreborg, K. H., Johansson, B., & Bodin, Ö. (2016). Systematic exploration of scenario spaces. *Foresight*, *18*(1), 59–75.

Coates, J. F. (1999). Getting at assumptions is troublesome. *Technological Forecasting and Social Change*, *62*(1), 97–99.

Collaborative, T. F. (2023). Insights into the accuracy of social scientists' forecasts of societal change. *Nature Human Behaviour*, *7*, 484–501.

Cordova-Pozo, K., & Rouwette, E. A. (2023). Types of scenario planning and their effectiveness: A review of reviews. *Futures*, *149*, Article 103153.

Fergnani, A., & Chermack, T. J. (2021). The resistance to scientific theory in futures and foresight, and what to do about it. *Futures & Foresight Science*, *3*(3–4), Article e61.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23.

Gertler, P., & Hofmann, B. (2018). Monetary facts revisited. *Journal of International Money and Finance*, *86*, 154–170.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Gordon, T. J. (2009). Cross-impact analysis. In J. C. Glenn, & T. J. Gordon (Eds.), *Futures research methodology version 3.0*. The Millennium Project.

Helmer, O. (1981). Reassessment of cross-impact analysis. *Futures*, *13*(5), 389–400.

Heugens, P. P., & van Oosterhout, J. (2001). To boldly go where no man has gone before: Integrating cognitive and physical features in scenario studies. *Futures*, *33*(10), 861–872.

Inayatullah, S. (2008). Six pillars: Futures thinking for transforming. *Foresight*, *10*(1), 4–21.

Jenkins, L. (1997). Selecting a variety of futures for scenario development. *Technological Forecasting and Social Change*, *55*(1), 15–20.

Kluyver, C. A. d., & Moskowitz, H. (1984). Assessing scenario probabilities via interactive goal programming. *Management Science*, *30*(3), 273–278.

Kowalski, K., Stagl, S., Madlener, R., & Omann, I. (2009). Sustainable energy futures: Methodological challenges in combining scenarios and participatory multi-criteria analysis. *European Journal of Operational Research*, *197*(3), 1063–1074.

Lawrence, M., O'Connor, M., & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*, *122*(1), 151–160.

Lewin, K. (1951). *Field theory in social science*. New York: Harper.

Lord, S., Helfgott, A., & Vervoort, J. M. (2016). Choosing diverse sets of plausible scenarios in multidimensional exploratory futures techniques. *Futures*, *77*, 11–27.

Loretan, M. (1997). *Generating market risk scenarios using principal components analysis: Methodological and practical considerations*: *Internal report*, Federal Reserve Board.

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science*, *65*(3), 621–649.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Novosyolov, A., & Satchkov, D. (2008). Global term structure modeling using principal component analysis. *Journal of Asset Management*, *9*, 49–60.

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, *73*(sup1), 69–81.

Petersen, J., & Steinmueller, K. (2009). Wild cards. In J. C. Glenn, & T. J. Gordon (Eds.), *Futures research methodology version 3.0*. The Millennium Project.

Phillips, A. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957. *Economica*, *25*(100), 283–299.

Prachowny, M. F. J. (1993). Okun's law: Theoretical foundations and revised estimates. *The Review of Economics and Statistics*, *75*(2), 331–336.

Ramirez, R., Mukherjee, M., Vezzoli, S., & Kramer, A. M. (2015). Scenarios as a scholarly methodology to produce "interesting research". *Futures*, *71*, 70–87.

Ritchey, T. (2006). Problem structuring using computer-aided morphological analysis. *Journal of the Operational Research Society*, *57*(7), 792–801.

Roponen, J., & Salo, A. (2023). A probabilistic cross-impact methodology for explorative scenario analysis. *Futures & Foresight Science*, Article e165.

Rowe, G., & Wright, G. (1999). The delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, *15*(4), 353–375.

Salo, A., Tosoni, E., Roponen, J., & Bunn, D. W. (2022). Using cross-impact analysis for probabilistic risk assessment. *Futures & Foresight Science*, *4*(2), Article e2103.

Schweizer, V. J. S., & Kriegler, E. (2012). Improving environmental change research with systematic techniques for qualitative scenarios. *Environmental Research Letters*, *7*(4), Article 044011.

Seeve, T., & Vilkkumaa, E. (2022). Identifying and visualizing a diverse set of plausible scenarios for strategic planning. *European Journal of Operational Research*, *298*(2), 596–610.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?*. Princeton: Princeton University Press.

Tetlock, P. E., & Gardner, D. M. (2015). *Superforecasting: The art and science of prediction*. New York: Crown.

Tetlock, P. E., Mellers, B., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and the quality of debate. *Current Directions in Psychological Science*, *23*(4), 290–295.

Tietje, O. (2005). Identification of a small reliable and efficient set of consistent scenarios. *European Journal of Operational Research*, *162*(2), 418–432.

van der Heijden, K. (2005). *Scenarios: The art of strategic conversation* (2nd ed.). Chichester: Wiley.

Vasara, A. (2020). *Theory of reflexive control: Origins, evolution and application in the framework of contemporary Russian military strategy*. Helsinki: National Defence University.

Wheatcroft, E. (2019). Interpreting the skill score form of forecast performance metrics. *International Journal of Forecasting*, *35*(2), 573–579.

Wright, G., Saunders, C., & Ayton, P. (1988). The consistency, coherence and calibration of holistic, decomposed and recomposed judgemental probability forecasts. *Journal of Forecasting*, *7*(3), 185–199.